Bachelor Thesis

An Interpretable Approach to Fake News Detection

Caio Brighenti

Date: April 25, 2020

Advisors: Prof. Michael Hay and William Cipolli

Technical Report: COSC-TR-2020

Department of Computer Science
Colgate University
Hamilton, New York

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris. …

Acknowledgments

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

. . . .

(Caio Brighenti)

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Our Contribution 2

1.3 Organization 2

2 Related Work **3**

2.1 Levels of Analysis 3

2.2 Feature Selection 4

2.3 Model Types 5

3 Methodology **7**

3.1 Dataset 7

3.2 Feature Engineering 8

3.3 Outlier Removal 14

3.4 Data Analysis and Modeling 14

4 Results **17**

4.1 Mood’s Median Test 17

4.2 Modeling 20

5 Conclusion **25**

5.1 Contributions 25

5.2 Future Work 26

1 Introduction

1.1 Motivation

Propaganda has long been a tool of political influence, but in recent years it has taken a new online form: fake news. Fake news, once a buzzword on the internet, is now at the center of global politics, one of the most common bigram in the lexicon of United States president Donald Trump. After the term gained prominence in Trump's presidential campaign in 2016, it exploded into public conciousness, earning the distinction of Webster-Collins' "Word of the Year" in 2017.[1] As the current presidential race unfolds, fake news has returned to the center of the conversation, with major social media companies facing scrutiny of their misinformation policy. This phenomenon is also not a distinctly American problem–investigate reporting both during and after the 2018 Brazilian president election demonstrated that more than 40% of right-wing viral news articles shared on the popular messaging service WhatsApp were fake news favoring the eventual winner, Jair Bolsonaro.[2]

Fake news is not only politically significant but also dangerously tempting. Studies have shown false content propagates faster through social media than real content.[3] Blatantly false or exaggerated rhetoric can even lead to violent action, as demonstrated by the "Pizzagate" incident in which a man stormed a D.C. pizzeria with an AR-15, having been convinced by false and unverified information that a pedophile ring operated out of the restaurant's basement.[4] Fake news can also be incredibly easy to create. In 2019, a group of researchers at the Allen Institute for Artificial Intelligence published text generation model able to produce fake news.[5] In a troubling conclusion, the researchers found that state-of-the-art fake news detection systems struggled more with identifying fake news produced by their systems than actual fake news.[6] Fake news is thus easy to create, spreads quickly, and is hard to detect, a dangerous combination making it a serious threat to civic society.

[1] [?]

[2] [?]

[3]

[4]

[5]

[6]

1.2 Our Contribution

Given the danger that fake news poses, machine learning and natural language processing researchers have devoted significant attention to the problem of fake news classification. However, previous attempts at fake news classification overwhelmingly rely on highly complex models suffering from the "black box" problem. As a result, these models lack interpretability and do not allow us to reach new conclusions about the nature of fake news. In order to begin closing this gap, this research adopts an interpretable approach, with the overall objective of a producing a fake news classification model with comparable accuracy to state of the art models without compromising interpretability.

The results of this paper make several contributions to the growing field of fake news detection. Firstly, it demonstrates that the results found in the few similar works that exist are not generalizable to the dataset used in this paper, finding few areas of agreement despite significant overlap in features used. More positively, this paper demonstrates the effectiveness of simple, interpretable models, producing a model performing similarly to deep learning approaches without sacrificing interpretability. Lastly, the results of this paper contribute to the growing understand of the textual characteristics of fake news, finding that the body of fake articles uses more exclamation points, prepositions, and auxiliary verbs, and uses more words per sentence, on average. Additionally, this paper finds that the titles of fake news articles also use mode exclamation marks than the titles of real articles, and use more words focusing on females and the human body.

1.3 Organization

Section 2 of this paper provides an extensive outline of previous work in fake news detection, describing how approaches tend to vary in levels of analysis, feature selection, and model type. Section 2 also demonstrates the need for interpretable approaches, making the case for the methodology pursued in this paper. Section 3 details the methodology employed in this paper, describing the dataset used, the feature engineering process, outlier removal, and the data analysis and modeling approach. Section 4 provides the results of the data analysis and modeling approach, including tables with significant results for Mood's Median Tests across all features, and plots showing the most impactful and important variables for fake news detection at the body-level and title-level. Finally, Section 5 summarizes the contributions of this paper, and offers several suggestions for further research informed by the results of the data analysis and modeling process.

2 Related Work

Since the 2016 U.S. presidential election, fake news has been a frequent topic of natural language processing research. There are countless examples of papers approaching fake news classification or closely related problems, but from slightly different angles. This section summarizes the prior work in fake news detection, clarifying the different categories of methods. In general, previous works in fake news detection differ in three major ways: 1) the scale of the predicted variable, the information used as features, and the type of model. With respect to scale, any fake news detection model falls into one of three levels of granularity: 1) claim level, 2) source level, and 3) article level. These levels of analysis describe the response variable being predicted.

2.1 Levels of Analysis

Claim level approaches attempt to determine whether a given claim, usually one or several sentences, is true or intentionally misleading. [1] Given that claim-level approaches must make a judgement based on only a short amount of text, researchers often adopt a fact-checking strategy This strategy, also known as "truth discovery," assumes that a sentence's claims can be gramatically isolated and checked against a database of established claims.[2] A natural application for claim-level models is social media, most commonly Twitter, where little is known about the author and only a very limited amount of text is available. However, claim-level approaches have serious limitations, often struggling with the complex sentences journalists or other writers typically employ.[3] Aditionally, they rely entirely on a complete knowledge base, which must be constantly expanded and updated, clearly a difficult task.

Source level approaches attempt to classify whether a speaker or entire news source consitently publish misinformation. The intuition behind these approaches is that speakers or sources that have published misinformation in the past are likely to continue to do so.

[1] Examples of claim-level approaches include ...

[2] strube p. 2

[3] strube 2

An example of a source-level approach is the popular browser extension "BS Detector,"[4] which classifies articles on a fine-grained scale of veracity by checking the source's status in a database of news sources and their reliability. A source's history of misinformation can also be used as a predictor in claim-level or article level approaches. Kirilin and Strube, for instance, create *Speaker2Credit*, a metric of speaker credibility, and show how it can improve the performance of fake news detection models when used as an input.[5]

Article level approaches have received the most attention in the work on fake news detection. This is logical, given that fake news tends to take the form of articles, peddling misinformation in the article text while posing as a legitimate source. Article-level approaches also benefit from a rich list of predictors to choose from, including not only the article's content but all relevant metadata. Kai Shu et. al, for instance, build an article-level model using linguistic and visual components of the article content, the social context around it—including information on the user that posted it, the post itself, responses to it, and the social network of the poster—, as well as spatiotemporal information capturing when and where the article and responses were to it were posted from.[6]

2.2 Feature Selection

The work of Shu et. al is an example of the overwhelming number predictors available to researchers working in fake news detection, resulting in a diversity of approaches with respect to feature selection. Melanie Tosik et. al, for instance, employ only hand-crafted features capturing the similarity between an article's title and text in a two-stage ensemble classifier modeling whether an article's body agrees with its headline.[7] Sonam Tripathi and Tripti Sharma demonstrate the effectiveness of parts of speech tagging—also known as grammatical tagging—in document classification problems, the general category of natural language processing that article-level fake news detection falls under.[8] Ramy Baly et al. employ a breadth of features to model factuality and bias of news sources, using features covering the content of articles, the source's Wikipedia and Twitter pages, the structure of the URL, and the source's web traffic. [9]

[4]

[5] Kirilin and Strube

[6] Shu et al

[7] Tosik et al

[8] tripathi

[9]

2.3 Model Types

Independent of level of analysis or choice of predictors, researchers overwhelmingly choose to use complex deep neural networks. Ajao et. al, for instance, use a "hybrid of convolutional neural networks and long-short term recurrent neural network models" to classify Tweets as true or false based on their text content.[10] The dominance of deep learning approaches is visible in an extensive survey on fake news detection done by Ray Oshikawa and Jing Qian. [11] The pair's section on machine learning models dedicates a total of three sentences to "Non-Neural Network Models," compared to seven paragraphs. focusing on neural networks.

While deep learning approaches can produce highly accurate models that consistently succeed in identifying misinformation, they also suffer from a lack of interpretability. This is often referred to as the "black-box" problem, meaning that the inputs and outputs of these models are perfectly clear, but the steps that the model takes to reach the output are completely invisible. This has been identified as a limitation of the the work on fake news detection thus far.[12] Oshikawa and Qian, at the conclusion of their extensive survey, declare that "we need more logical explanation for fake news characteristics," highlighting the need for models that can teach us something about fake news.

Approaches that focus on interpretability are rare, but do exist. From the deep learning perspective, Nicole O'Brien et al. employ post-hoc variable importance to their text-based deep learning model, identifying the words that are most predictive of fake and real news.[13] Their approach, however, does not interpret the results, but instead merely demonstrates the feasiblity of the technique. Furthermore, this method reveals only information about *specific* words, as opposed to *types* of words. More applicable is the work of researchers who both use features that describe the semantic properties of the text in general, use interpretable models, and extensively document their results. The best examples of this type of work are the works of Benajmin Horne et al. and Mauricio Gruppi et al., both of which employ features capturing the complexity, style, and psychology of fake news, and display precisely how fake and real news differs in each of the variables used.[14]

[10] ajao et al

[11] Oshikawa

[12] shu, O'Brien

[13] O'Brien et al

[14] Horne, Gruppi

3 Methodology

This paper seeks to contribute to the small literature of interpretable fake news detection, by following the methodology of Horne et al. and Gruppi et al., leveraging features that describe textual properties of fake news, applying non-neural network models and focusing heavily on interpretation of results. This section details the precise methodology employed, discussing the dataset used, the feature engineering process, outlier removal, and models applied.

3.1 Dataset

There are many datasets freely available for fake news detection, but many suffer quality limitations. Oshikawa and Qiang outline 12 requirements for a quality fake news dataset, expanding a 9-point list originally by Rubin et al.: 1. Availability of both truthful and deceptive instances; 2. Digital textual format acessibility; 3. Verifiability of ground truth; 4. Homogeneity in lengths; 5. Homogeneity of writing mattes; 6. Predefined timeframe; 7. The manner of news delivery; 8. Pragmatic concerns; 9. Language and culture; 10. Easy to create from raw data; 11. Fine-grained truthfulness; 12. Various sources or publishers.[1] While no dataset currently exists that meets all 12 criteria, the article-level *FakeNewsNet* (FNN) dataset meets most.[2]

FNN is an article-level dataset that includes the title and body of each article (2), each of which have been profesionally fact-checked and labeled as true or false by Politifact or Gossicop(1, 3). FNN provides both political and celebrity news articles, but this paper chooses to use only the political articles in order to maintain a roughly consistent corpus (4,5,7). These articles are all in English, largely center around American politics, and come from a variety of sources (9, 12). Finally, there is little work needed to obtain the dataset, as FNN provides an API to quickly obtain the body and title of each article (10). The biggest limitation for FNN is that it lacks a fine-grained scale of truth, labeling only as binary

[1] Oshikawa, Rubin

[2] FNN

true/false (11). However, given that it meets the majority of the criteria, and contains 726 observations, it is overall a good fit for this paper.

3.2 Feature Engineering

While FNN includes a host of metadata on each article, this paper utilizes only features engineered from the text and titles of each article. The objective of this paper is to reach new conclusions about the content of fake news, making certain metadata irrelevant or problematic. For instance, the usage of website traffic to gauge veracity may increase accuracy,[3] but measures nothing about the actual content of the article. It should be obvious that websites with high traffic are perfectly capable of producing misinformation. Aditionally, equating established sources with reliable sources can be problematic, failing to hold mainstream media accountable and preventing the growth of new, quality publishers.

The choice of using features capturing exclusively textual properties is motivated by a belief that the problem of fake news cannot be solved with deep learning models automatically policing all the content on the internet. A widespread system like this would have incredible power, and could easily become a force for opression and misinformation with biased data or improper use. Aditionally, people are unlikely to be convinced that something they believe to be true is misinformation just because a browser extension tells them it is. Fake news is a social problem, and will only be fixed with widespread education on how to identify fake news. To achieve this, this paper treats fake news detection as a learning opportunity, focusing solely on identifying the textual properties of text that might suggest an article is misinformation.

This paper uses the FNN dataset to obtain the title and body of articles as well as a true/false label, then leverages a series of natural language processing tools to engineer features describing the text. Each feature is calculated for both the body and the title seperately. Each feature falls under one of four categories: 1) complexity metrics, 2) summary, grammatical, and psychological metrics from the Linguistic Inquiry and Word Count engine,[4] 3) parts-of-speech tagging, and 4) named entity recognition. Each category represents a different method or tool for feature engineering. Each feature engineered is displayed in the tables below.[5]

The complexity metrics are calculated in three different ways. Several of these metrics are indexes of textual complexity calculated using the "'quanteda"' package in R. Others are

[3] forget which one does this

[4]

[5] And described in detail in Appendix A

variables describing the structure of verb-phrase and noun-phrase trees obtained for each sentence using the Stanford CoreNLP constituency parser. The final complexity metric is a manually computed type-token ratio, where types are all the unique words in a document and tokens are the total words in that document, capturing the diversity of the vocabulary used. As the tree depth measures capture the structure throughout a document, these features do not make sense for a single-sentence title, and are thus not included in the tests and models at the title-level.

Table 3.1: Complexity Metrics

Variable	Description
len	Document length
mu_sentence	Mean number of sentences
mu_verb_phrase	Mean depth of verb-phrase trees
mu_noun_phrase	Mean depth of noun-phrase trees
sd_sentence	Standard deviation of number of sentences
sd_verb_phrase	Standard deviation of depth of verb-phrase trees
sd_noun_phrase	Standard deviation of depth of noun-phrase trees
iqr_sentence	Interquantile range of number of sentences
iqr_verb_phrase	Interquantile range of depth of verb-phrase trees
iqr_noun_phrase	Interquantile range of depth of verb-phrase trees
num_verb_phrase	Number of verb-phrase trees
swc	Mean sentence word count
wlen	Mean word length
types	Number of unique words
tokens	number of total words
TTR	Type-token ration
FOG	Gunning's Fog Index
SMOG	Simple Measure of Gobbledygook
FK	Flesch-Kincaid Readability Score
CL	Coleman-Liau Index
ARI	Automated Readability Index

The LIWC features comprise a diverse range of different metrics, many capturing the different psychological components within the text such as cognitive processeses, or core drives and needs.[6]. Some LIWC variables capture simpler properties, such as the frequency

[6] For a full description of each of these variables, consult Appendix A or the LIWC Operator's Manual

of informal speech or specific punctuation marks. As LIWC's metrics are heavily dependent on counting words from dictionaries, each metric is normalized per 100 words, allowing them to be compared across differing document sizes.

Table 3.2: LIWC Metrics

Variable	Description
WC	Word count
Analytic	Words reflecting formal, logical, and hierarchical thinking
Clout	Words suggesting author is speaking from a position of authority
Authentic	Words associated with a more honest, personal, and disclosing text
Tone	Words associated with positive, upbeat style
WPS	Words per sentence
Sixltr	Number of six+ letter words
Dic	unsure
function	Function words
pronoun	Pronouns
ppron	Personal pronouns
i	1st person singular
we	1st person plural
you	2nd person
shehe	3rd person singular
they	3rd person plural
ipron	Impersonal pronoun
article	Articles
prep	Prepositions
auxverb	Auxiliary verbs
adverb	Common adverbs
conj	Conjuctions
negate	Negations
verb	Regular verbs
adj	Adjectives
compare	Comparatives
interrog	Interrogatives
number	Numbers

available at

quant	Quantifiers
affect	Affect words
posemo	Positive emotions
negemo	Negative emotions
anx	Anxiety
anger	Anger
sad	Sad
social	Social words
family	Family
friend	Friends
female	Female referents
male	Male referents
cogproc	Cognitive processes
insight	Insight
cause	Cause
discrep	Discrepancies
tentat	Tentativeness
certain	Certainty
differ	Differentiation
percept	Perceptual processes
see	Seeing
hear	Hearing
feel	Feeling
bio	Biological processes
body	Body
health	Health/illness
sexual	Sexuality
ingest	Ingesting
drives	Core drives
affiliation	Affiliation
achieve	Achievement
power	Power
reward	Reward focus
risk	Risk/prevention focus
focuspast	Past focus
focuspresent	Present focus

focusfuture	Future focus
relativ	Relativity
motion	Motion
space	Space
time	Time
work	Work
leisure	Leisure
home	Home
money	Money
relig	Religion
death	Death
informal	Informal speech
swear	Swear words
netspeak	Netspeak
assent	Assent
nonflu	Nonfluencies
filler	Fillers
AllPunc	All punctuation
Period	Periods
Comma	Commas
Colon	Colons
SemiC	Semicolons
QMark	Question marks
Exclam	Exclamation marks
Dash	Dashes
Quote	Quotes
Apostro	Apostrophes
Parenth	Parentheses (pairs)
OtherP	Other punctuation

All features falling under the third and fourth categories were obtained using the Stanford CoreNLP toolkit. Specifically, the grammatical incidence variables were obtained using the CoreNLP parts-of-speech tagger, which counts the frequency of types of words (for instance, verbs) within a document. As before, these metrics were normalized to account for differing document lengths. Finally, the named entity recognition feature is obtained using the CoreNLP named entity recognition annotator, which simply counts the total number of words that refer to named proper nouns. As these metrics are computed by summing appearances throughout a document, they were normalized by document length to correspond to per 100 words, similarly to the LIWC metrics.

Table 3.3: POS Metrics

Variable	Description
CC	Coordinating conjunctions
CD	Cardinal numeral
DT	Determiner
EX	Existential
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Ordinal number
JJR	Comparative adjective
JJS	Superlative adjective
LS	List item marker
MD	Model verb
NN	Noun, singular or mass
NNS	Plural noun
NNP	Singular proper noun
NNPS	Plural proper noun
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP.	Possessive pronoun
RB	Adverb
RBR	Comparative adverb
RBS	Superlative adverb
RP	Particle
SYM	Symbol
TO	To
UH	Exclamation/interjection
VB	Verb, base form
VBD	Past tense verb
VBG	Present participle
VBN	Past participle

VBP	Present tense verb, other than 3rd person singular
VBZ	Present tense verb, 3rd person singular
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

3.3 Outlier Removal

The complete training set thus consisted of 726 observations with 152 features each, constructed using the body and title of each article obtained by the FNN API, which queries the stored article URLs. However, given that web pages often change structure, move to different addresses, or are removed from the internet entirely, many of the entries have changed since the dataset was initially compiled, and are now unavaiable or in an incorrect format and must be removed from the dataset. To identify outliers, a baseline logistic regression model using all features was fit in order to identify overly influential observations with Cook's Distance four times larger than the mean. This approach cannot perfectly identify all outliers, but suggests that these observations are *potential* outliers. Each of the 575 potential outliers were manually inspected and labeled as either false positives or true outliers. Out of the 575 potential outliers, 422 were true outliers, 132 of which were Politifact articles, bringing the number of observations in the training set down to 594.

3.4 Data Analysis and Modeling

After having completed all feature engineering and outlier removal, exploratory data analysis was performed to identify group differences in each predictor between true and false articles. This approach reflects the work Horne et al. and Gruppi et al., allowing for the comparison of results between shared predictors. However, instead of applying an ANOVA test to each predictor, a Mood's Median hypothesis test was performed using the "'RVAdeMemoire"' package[7] as many of the predictors did not meet the normality assumption required by ANOVA and other traditional tests. Aditionally, estimates and 95% confidence intervals for each predictor accross both groups were calculated using bootstrapping 1,000 samples.[8]

[7]

[8] The results of the Mood's Median tests are displayed in the next section, but the confidence intervals are only shown in Appendix A.

To verify that power of the Mood's Median Test was not inflated with the sample size in this study, an experiment was conducted to verify the false positive rate of the test at this sample size. This experiment consisted of generating 1000 samples of 594 observations of random normally distributed data, with each observation having class of 0 or 1, to which a Mood's Median Test was applied. If the test performs properly, only 50 out of 1000 samples should result in p-values below 0.05. This experiment resulted in 47 'significant' samples, performing nearly exactly as expected for a test with 95% confidence. This confirms that the results of the Mood's Median Tests are not likely to be products of a large sample size.

Given the extensive number of predictors and the multicolinearity between many of them, a binomial Lasso regression—also known as logistic regression with L1 regularization—was applied to avoid overfitting by creating a more parsimonious model. Aditionally, to avoid skewed accuracy results due to class imbalance, the dataset was upsampled to have an equal proportion of negative and positive labels. The upsampling increased the number of negative articles to have a 374-374 split, as opposed to the original 374-220 split. After upsampling, a Lasso model was fit as the extensive number of predictors, many of which are correlated, would create a model suffering from overfitting and multicolinearity. The regularization parameter λ was selected through cross validation. The final λ selected was not the one that resulted in the lowest mean squared error, but rather the largest λ within 1 standard error from the 'ideal' λ. This was done in order to produce a more parsimonious model, due to the large number of predictors.

After fitting the Lasso model, the features reduced to zero were removed and a logistic regression model using only the preserved features.This is the model used for final interpretation of results. Aditionally, seperate models are fit for the body and title of articles. An approached focused entirely on predictive accuracy would likely instead create a two-stage ensemble model, but as the overall objective is interpretation, maintaining the two seperate is preferable as it allows for better interpretation of the results.

The results of the models are shown in the form of plots capturing the most important and impactful variables. Variable importance is measured using the "caret"[9] package in R. The variable importance metric used is the AUC of each feature when used in a univariate model predicting the class in question. This gives a measure of each feature's individual predictive strengthm with a baseline of 0.5. The feature with the highest variable importance score has the highest individual predictive power across the entire dataset. Variable impact is measured using the coefficients of the final binomial logistic regression model. While 'important' variables have high predictive power across the entire dataset, 'impactful' ones have the highest effect at the observation level when taking on a value significantly higher or lower than the mean for that feature.

[9]

4 Results

This section summarizes the results of the analysis, starting with the pairwise Mood's Median tests followed by the results of the modeling process. Results are compared with the work of Horne et al. and Gruppi et al. to highlight overlaps and disagreements between their results and the results of this paper.

4.1 Mood's Median Test

For each feature described in **??**, a Mood's Median Test was used to test for differences between real and fake news. The table below includes the result of these tests, excluding tests with resulting p-values of > 0.05. For each predictor, the p-value of the test is given, along with a comparison of the groupwise medians to indicate which group is higher. Furthermore, as many of the features used in this paper are also used in O'Brien et al. and Gruppi et al., we indicate, when available, whether the results of these tests agree with the results found in both papers. As both papers First, a table including results for the body of articles is included, then for the titles.

Table 4.1: Mood's Median Test Results for Article Body

Variable	Result	p-value	Gruppi et al.	Obrien et al.
mu_sentence	Real > Fake	< 0.001	-	Disagree
mu_verb_phrase	Real > Fake	< 0.001	-	-
num_verb_phrase	Fake > Real	< 0.05	-	-
swc	Real > Fake	< 0.001	-	-
types	Fake > Real	< 0.05	-	-
tokens	Fake > Real	< 0.001	-	-
TTR	Real > Fake	< 0.001	-	Disagree
FOG	Real > Fake	< 0.001	-	-
SMOG	Real > Fake	< 0.001	Agree	-
FK	Real > Fake	< 0.001	Disagree	Agree

CL	Fake > Real	< 0.01	-	-
ARI	Real > Fake	< 0.001	-	-
WC	Fake > Real	< 0.01	Disagree	-
Tone	Fake > Real	< 0.001	-	-
WPS	Real > Fake	< 0.001	Agree	-
shehe	Real > Fake	< 0.001	Agree	Disagree
article	Real > Fake	< 0.001	-	-
prep	Real > Fake	< 0.001	-	-
auxverb	Real > Fake	< 0.05	Agree	-
conj	Fake > Real	< 0.001	-	-
posemo	Fake > Real	< 0.001	-	-
negemo	Real > Fake	< 0.001	-	-
anger	Real > Fake	< 0.001	-	-
male	Real > Fake	< 0.001	-	-
cogproc	Fake > Real	< 0.001	-	-
insight	Real > Fake	< 0.05	-	-
discrep	Fake > Real	< 0.05	-	-
differ	Fake > Real	< 0.01	-	-
percept	Real > Fake	< 0.05	-	-
see	Real > Fake	< 0.001	-	-
hear	Real > Fake	< 0.001	-	-
feel	Real > Fake	< 0.01	-	-
bio	Real > Fake	< 0.001	-	-
body	Real > Fake	< 0.01	-	-
health	Real > Fake	< 0.05	-	-
drives	Fake > Real	< 0.001	-	-
affiliation	Fake > Real	< 0.001	-	-
achieve	Fake > Real	< 0.001	-	-
focuspast	Real > Fake	< 0.001	-	-
focuspresent	Fake > Real	< 0.001	-	-
relativ	Real > Fake	< 0.01	-	-
time	Real > Fake	< 0.001	-	-
work	Fake > Real	< 0.05	-	-
leisure	Real > Fake	< 0.001	-	-
money	Fake > Real	< 0.001	-	-
AllPunc	Fake > Real	< 0.01	-	-

Period	Fake > Real	< 0.001	-	-
Dash	Fake > Real	< 0.01	-	-
Quote	Real > Fake	< 0.001	-	Agree
Apostro	Real > Fake	< 0.05	-	-
Parenth	Fake > Real	< 0.01	-	-
CC	Fake > Real	< 0.01	-	-
DT	Fake > Real	< 0.05	-	Disagree
IN	Fake > Real	< 0.05	-	-
JJ	Fake > Real	< 0.01	-	-
MD	Fake > Real	< 0.05	-	-
NN	Fake > Real	< 0.001	-	Disagree
NNS	Fake > Real	< 0.001	-	-
NNP	Fake > Real	< 0.05	-	-
NNPS	Fake > Real	< 0.001	-	-
TO	Fake > Real	< 0.001	-	-
VB	Fake > Real	< 0.05	-	-
VBN	Fake > Real	< 0.01	-	-
VBP	Fake > Real	< 0.05	-	-
VBZ	Fake > Real	< 0.01	-	-
NER	Real > Fake	< 0.001	-	-
len	Fake > Real	< 0.001	-	-

Despite significant overlap in features used with two comparable papers, particularly O'Brien et al., very few features significant in this paper's tests were also significant in the other two studies. Furthermore, when features were significant in both cases, they frequently disagreed, in many cases even demonstrating agreement for one study and disagreement in another. The only clear agreements were the SMOG complexity score, mean words per sentence, auxiliary verbs per 100 words, and quotes per 100 words with respect to the body of articles.

Next, the results for the same tests and features with respect to the title of articles is shown below. As Gruppi et al. do not model at the title level, only comparisons with O'Brien et al. are included.

Table 4.2: Mood's Median Test Results for Article Title

Variable	Result	p-value	Obrien et al.

mu_sentence	Real > Fake	< 0.001	-
mu_verb_phrase	Real > Fake	< 0.001	-
mu_noun_phrase	Real > Fake	< 0.05	-
sd_noun_phrase	Real > Fake	< 0.05	-
num_verb_phrase	Real > Fake	< 0.001	-
types	Real > Fake	< 0.001	-
tokens	Real > Fake	< 0.001	-
TTR	Fake > Real	< 0.001	-
SMOG	Real > Fake	< 0.001	-
FK	Real > Fake	< 0.05	Agree
CL	Fake > Real	< 0.01	-
ARI	Real > Fake	< 0.05	-
WC	Real > Fake	< 0.001	-
WPS	Real > Fake	< 0.001	Disagree
function.	Real > Fake	< 0.001	-
prep	Real > Fake	< 0.05	-
verb	Real > Fake	< 0.001	-
social	Real > Fake	< 0.001	-
space	Real > Fake	< 0.05	-
NNP	Real > Fake	< 0.001	Disagree
len	Real > Fake	< 0.001	-

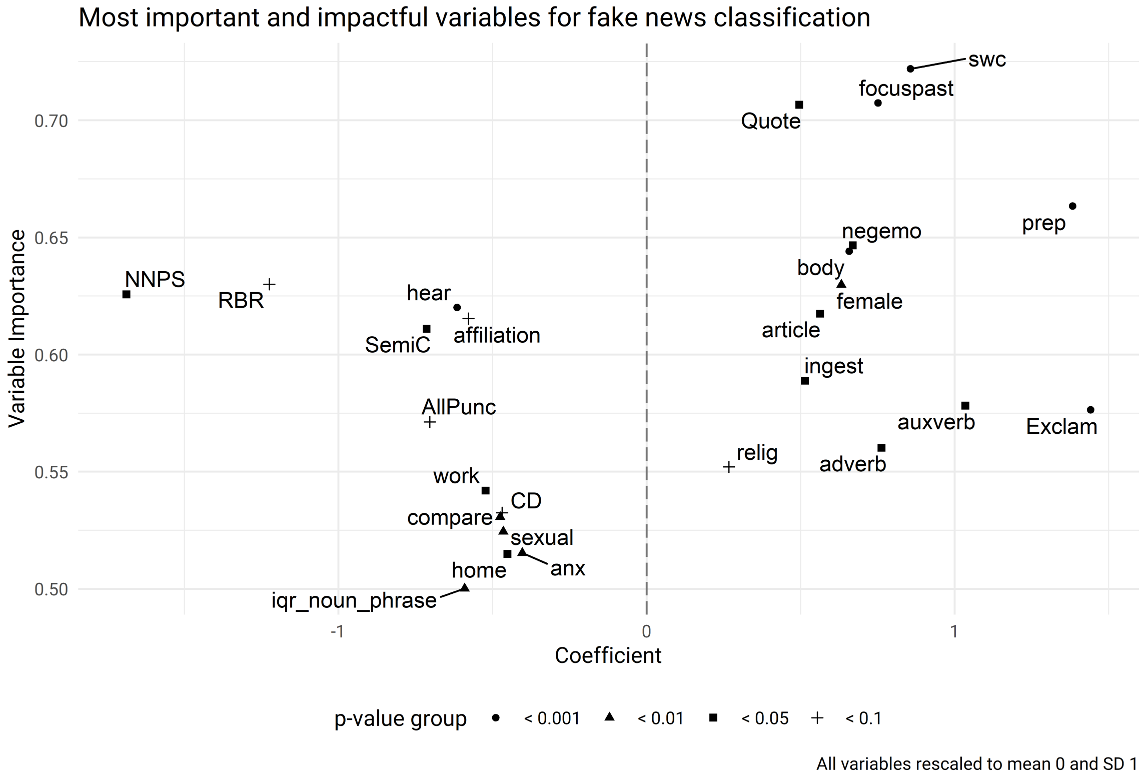
Similarly to the tests describing the body of articles, there was little overlap between significant results in this paper and in the work of O'Brien et al, with only three features significant in both. Additionally, the pattern of disagreement continued, with two out of the three overlapping results in disagreement. The only result shared between this paper and O'Brien et al. was that the FK complexity score is higher in real articles than fake articles.

4.2 Modeling

While the results of the Mood's Median Tests demonstrate where real and fake articles differ, on median, at the level of each feature, it does not provide a meaningful way to predict the class of an article based on these textual features. While we expect features with significant tests to be quality predictors, this is not always the case. This section thus highlights the most important and impactful predictors of fake news, utilizing the coefficients and AUC-based variable importance. The overall predictive accuracy of each model is also described and compared to prior work with the FNN dataset, to demonstrate the efficacy of this paper's model.

The final body-level model included X features after the Lasso regression shrunk Y features to 0. This model performs quite well, with an AUC of 0.85, and a accuracy of 76% relative to a 50% baseline, as well as sensitivity and specificity of 0.74 and 0.78 with a cutoff of 0.5. This is close to BLABLABLABLA COMPARABLE RESULTS GO IN HERE!

The most impactful and important features for the body-level model are shown in the following plot, with the coefficient magnitude shown on the x-axis and the AUC-based variable importance on the y-axis. Note that the baseline importance for a variable with no predictive power is 0.5, hence why the y-axis does not start at 0. Features with large variable importance scores have the most individual predictive power, while features with large coefficients having the largest effect on individual observations when these features take on values significantly away from the mean. Negative coefficients represent features indicative of real news, while positive ones are indicative of fake news. Note that only features with p-values of >0.1 are shown.

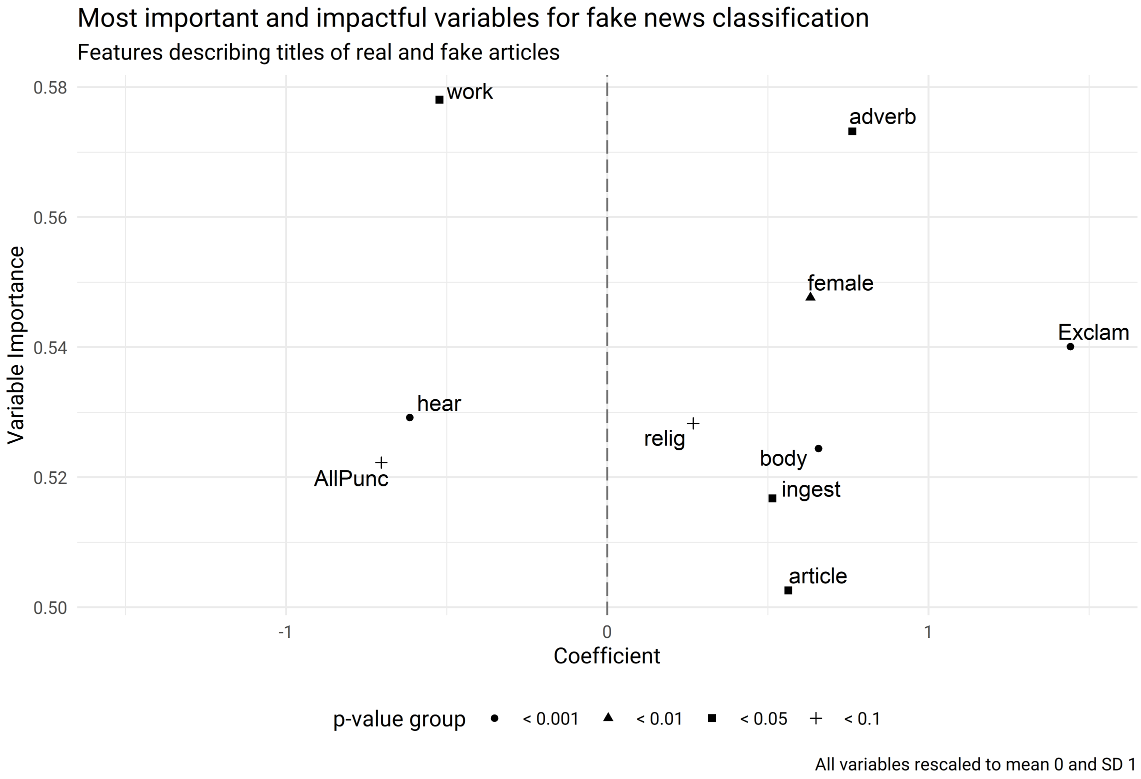

Most important and impactful variables for fake news classification

All variables rescaled to mean 0 and SD 1

The above plot demonstrates that the most impactful predictors of fake news at the body-level are the usage of prepositions, auxiliary verbs, and exclamations. Predictors indicating news is real are less numerous, with the most impactful being the usage of plural proper nouns. As for most important features, the usage of quotes and words focusing on the past above the mean are good individual predictors of fake news across the entire dataset, as well as the length of sentences.

The final title-level included X features after the Lasso regression shrunk Y features to 0. This model performs similarly to the body-level model, with an AUC of 0.835, and an accuracy of 77% relative to a baseline of 50%, as well as a sensitivity and specificity of 0.823 and 0.729 respectively, using a cutoff of 0.36 selecting using ROC curve analysis. Again, this accuracy is similar to prior work, comparable with the work of BLANK AND BLANK.

The same feature importance and impact plot created for the title-level model, and is shown below. The interpretation for this plot is the same as before. As before, only features with p-values below 0.1 are shown.



Similarly to the results of the Mood's Median Tests, there are fewer significant features at the title-level than at the body-level. The most impactful predictor of fake news is again the usage of exclamation marks, suggesting it is a consistent point of difference between real and fake news. The most impactful predictors of real news are words relating to work, the perceptual process of hearing, and the usage of punctuation in general, though the latter is only marginally significant. The most important individual predictors are words relating to work, suggesting an article is real, and the usage of adverbs, suggesting an article is fake.

5 Conclusion

This section summarizes the contributions of this paper, describing the effectiveness of an interpretable text-only model, discussing the implications of the divergence in results to comparable papers, outlining the biggest takeways with respect to identifying fake news, and finally offering suggestions for future work in this area.

5.1 Contributions

This paper set out to build an interpretable model using only features describing the textual properties of fake news. Aditionally, one objective was to demonstrate that such a model can perform close to deep learning approaches, while simultaneously preserving the learning opportunity offered by interpreting results and human-understandable features. Given that the accuracy of both models were similar to that of more advanced models using the same dataset (FNN), it is clear that one need not forgo accuracy entirely when focusing on interpretability.

A significant contribution of this paper is the divergence from prior work following a similar methodology. Given that so few works exist deeply engaging with features describing the text of fake news, it is important to reproduce and verify the little work that has been done. Unfortunately, few of the results in this paper agree with the prior work of O'Brien et al. and Gruppi et al. This suggests that while fake news detection using text factors is certainly possible within a dataset, it may not generalize well to other datasets.

Considering only the observed results of this paper, several text features make for good predictors of fake news. Fake news uses more prepositions and auxiliary verbs, focuses more on the past, tends to have longer sentences, and uses more exclamation marks, on average, in the body of articles. At the title level, this paper finds less significant distinctions between fake and real articles, though the model is still quite accurate. Titles of fake news articles tend to use more exclamation points, adverbs, words referring to females, and words referring to the human body than the titles of real news.

5.2 Future Work

The results of this paper suggest several different viable research paths for fake news detection. Firstly, it is obvious more works following the methodology of this paper, Gruppi et al., and O'Brien et al. Given the significant divergence between the results across the three papers, there is a serious need for further work to identify which properties of fake news are truly significant and not spurious or a product of the particular dataset. Additionally, further work of this type should apply the final model to a completely different dataset, and not simply an out-of-sample test set belonging to the same dataset.

The results of this paper also demonstrate that the gap in interpretable studies is unjustified, given the performance of a simple logistic regression model. As the overreliance on black box models has been identified as a significant limitation of prior work in this field, it would seem particularly prudent for future researchers to follow this direction as opposed to constantly racing for the next state-of-the-art model.

Finally, while an interpretable model can serve as a learning opportunity regarding fake news, offering specific textual properties distinguishing real and fake news, a computer science researcher is not qualified to fully interpret these properties. The overall research objective of identifying relevant textual properties of fake news and using them to better understand the phenomenon in general can only be done in collaboration with researchers from other fields, such as linguists and psychologists. Given the threat that fake news poses, this type of interdisciplinary work is critical, and should be pursued immediately. If understanding fake news is not treated as a serious research objective, the democratic process of the upcoming U.S. presidential election is in significant risk of being compromised.